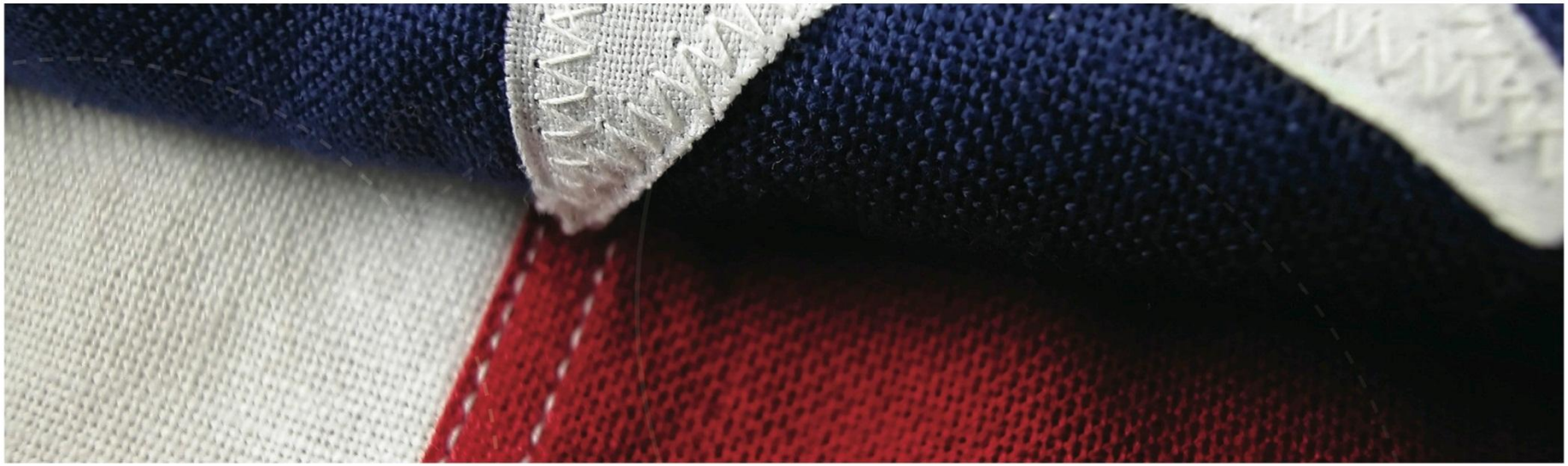


Using Different Intra-class Correlations to Assess Inter-rater Reliability + Inter-rater Agreement: Example of Organizational Readiness to Change from Three Implementation Studies

Fifth Annual NIH Conference on the Science of Dissemination and Implementation
March 19, 2012

Christian D. Helfrich, MPH, PhD – VA Northwest HSR&D Center of Excellence
Dean Blevins, PhD – Centers for Disease Control and Prevention
P. Adam Kelly, PhD – Southeast Louisiana Veterans Health Care Network
Ina M. Gylys-Colwell, MS – VA Northwest HSR&D Center of Excellence
Patricia M. Dubbert, PhD – South Central VA Mental Illness Research, Education and Clinical Center

VA QUERI
Quality Enhancement Research Initiative

VA HEALTH CARE Defining EXCELLENCE in the 21st Century

Background: Importance of agreement in survey measures of latent organizational constructs

- Often interested in latent organizational constructs
  - E.g., workplace climate for patient safety
- Assumes shared experiences / shared interpretation of experiences
- Difficult to directly observe and measure
- Measure via surveys of multiple individuals
  - Aggregated to unit measure (e.g., a team, facility)

Background: Importance of agreement in survey measures of latent organizational constructs

- Prior to aggregating, need to assess level of agreement
- Two underlying issues
  - Reliability of the measure
  - Construct validity

Background:
Inter-rater reliability + inter-rater agreement

- Two distinct concepts of agreement within groups (LeBreton et Senter 2008) :
  1. Inter-rater agreement (IRA): The reliability of the respondents' scores in terms of ***absolute scores*** of groups
  2. Inter-rater reliability (IRR): The reliability of respondents' scores in terms of ***rankings*** of groups
- We are interested in both.
- Remainder of slides use inter-rater reliability to refer to both.

Background: Importance of agreement in survey measures of latent organizational constructs

- Inter-rater reliability different than internal-consistency reliability
- Different than in inter-rater agreement in clinical or qualitative methods contexts
  - Reliability of diagnosis, Cohen's kappa
  - Repeated measures reliability

Background:
Intra-Class Correlation Coefficients (ICCs)

- Intra-class correlation coefficients (ICC) most common IRA+IRR measure
- Six different types of ICC – study design x unit of reliability (Shrout et Fleiss 1979)
- Two applicable for designs where respondents don't overlap sites
  - ICC(1) – reliability of individual-level score as representation of group
  - ICC(2) - reliability of group-mean score to distinguish among groups

Background: ICC(1)

- ICC(1) is between-group variance (MS_R) minus within-group variance (MS_W) over total variance adjusted for number of respondents per site (n_K).
  - $ICC(1) = \frac{MS_R - MS_W}{MS_R + (n_K - 1)MS_W}$
- Indicates the reliability of the individual respondents' scores.
- Can be interpreted as an effect size.
- Conventional Threshold >=.08-.20 (LeBreton et Senter 2008)

Background: ICC(2)

- ICC(2) is between-group variance (MS_R) minus within-group variance (MS_W) over Between-group variance (MS_R).
  - $ICC(2) = \frac{MS_R - MS_W}{MS_R}$
- The reliability or consistency of the group mean score (i.e., how much will the score shift by virtue of who responds).
- Conventional threshold >= .70 (LeBreton et Senter 2008)

Background: ICC(2)

- ICC(2) sensitive to sample size
- Number of respondents/site needed to obtain a given ICC(2) can be calculated from Spearman-Brown formula based on ICC(1) (Shrout et Fleiss, 1979).

$$n_k = \frac{ICC2(1 - ICC1)}{ICC1(1 - ICC2)}$$

Background
Organizational Readiness to Change Assessment

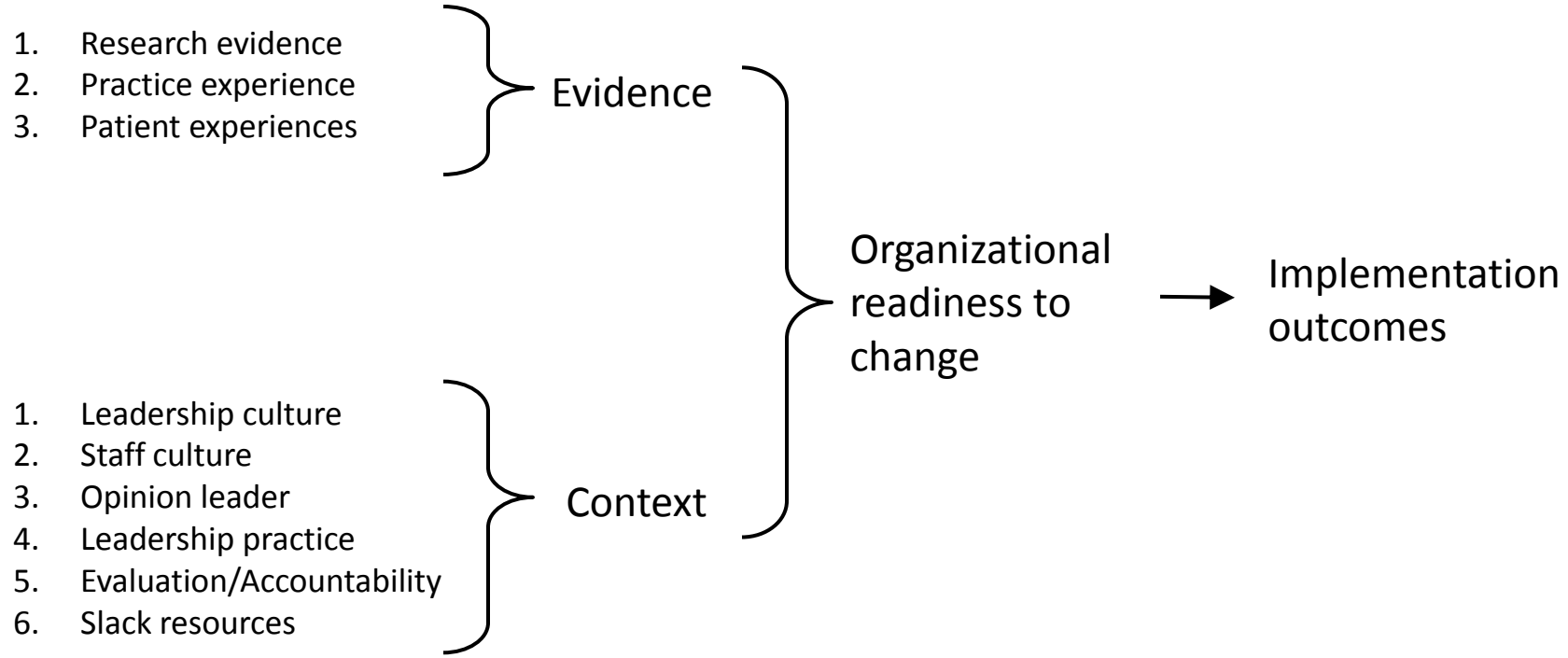
- Survey intended to be fielded at baseline of implementation study/project, i.e., implementation of a specific evidence-base practice.
- Prognostic and diagnostic uses
- 77-items, 19 subscales, 3 scales: Evidence, Context and Facilitation.
  - This analysis focused on Evidence and Context because of stage of implementation.
- Fielded among clinicians & staff involved in implementation.

Background
Organizational Readiness to Change Assessment

Development

- Resulted from QI projects by the Ischemic Heart Disease Quality Enhancement Research Initiative (IHD QUERI).
- Organized according to the Promoting Action on Research Implementation in Health Services (PARIHS) framework.
- Initially fielded in three quality improvement projects.

Background:
Organizational Readiness to Change Assessment



Background
Organizational Readiness to Change Assessment

Scoring

- Items are statements.
  - E.g., “The proposed practice changes or guideline implementation are supported by randomized controlled trials or other scientific evidence.”
- 5-point Likert scale, strongly disagree (1) - strongly agree (5).
- Subscales comprise 3-6 items
- Higher scores hypothesized to favor implementation.

Specific Aims

1. Test the inter-rater reliability and inter-rater agreement of two ORCA scales with ICC1 and ICC2, against conventional thresholds of reliability.
2. Determine minimum number of respondents per site needed to obtain a reliable site-level ORCA score.

These analyses are part of a broader study of the psychometrics of the ORCA (Helfrich et al 2011)

Methods

- Aim 1: One-way ANOVA (loneway) with STATA (v. 11)
  - Comparing study site means on ORCA scales
  - Separate models for Evidence, Context
- ICC1 compared to threshold of .08
- ICC2 threshold of .70 (LeBreton & Senter 2008, James 1982)
- Aim 2: Spearman-Brown formula used to estimate number of respondents needed based on observed ICC1 (Shrout & Fleiss 1979)

Findings: Partner studies

		Respondent n (Respondents/site)		
Partner study	Site n	Evidence	Context	Study response rate
1. Implementing cognitive behavioral therapy as first line Tx for depression in primary care	18	26 (1.4)	26 (1.4)	65%
2. Increasing enrollment in VA personnel health record among Veterans with spinal cord injury	2	9 (4.5)	19 (9.5)	96%
3. Implementation Hep-C screening and treatment in substance use disorder clinics	21	60 (2.9)	60 (2.9)	Unknown
Total	**41**	**95 (2.3)**	**105 (2.7)**	

Findings: Aim 1
Does the ORCA meet inter-rater reliability thresholds?

	ICC1	95% CI	ICC2	Prob > F
Evidence	.32	.07-.57	.52	.006
Context	.27	.03-.50	.48	.01

Findings: Aim 2
How many respondents per site would we need?

		Desired level of reliability (ICC2)			
	Scale	.60	.70	.80	.90
Number of respondents needed per site	Evidence	3.2	5.0	8.5	19.1
	Context	4.1	6.3	10.8	24.3

Discussion

- ICC1 results support the construct validity of the instrument as a measure of an organizational-level construct.
- ICC2 results indicate mean scores could not be reliably estimated at the organizational level.
  - Much larger numbers of respondents per site are needed to obtain reliable site-level measures.
    - Minimum 9 observations/site for Evidence
    - Minimum 11 observations/site for Context

Implications

- May not be possible to obtain site-level score
- May not be appropriate to obtain site-level score
  - Multiple views may be critical; defining subgroups
  - Agreement or dispersion may be important indicator in own right

Limitations

- Internal validity
  - We don't have information on respondent characteristics.
    - Notably supervisory level; cannot tell how much variation attributable to position.
- External validity
  - The three studies are all within the Veterans Health Administration, may not generalize to other populations.
- Construct validity
  - Work left to do on defining readiness, e.g., over time, among subgroups

Future Directions

- Predictive validation - implementation effectiveness
- Explore association of site-level variance with implementation effectiveness
- Compare ORCA measures at baseline and follow-up to see if ICCs change

Acknowledgements

- Psychometric validation of an organizational readiness-to-change scale (VA RRP 07-280)
- Anne E. Sales
- Nancy Sharp
- Yu-Fang Li

- Predicting implementation from organizational readiness to change (VA IIR 09-067)
- Anne E. Sales
- Hildi Hagedorn
- Timothy Hogan
- Rick Owen
- Jeffrey Smith

References

Hagedorn, H. J. and P. W. Heideman (2010). "The relationship between baseline Organizational Readiness to Change Assessment subscale scores and implementation of hepatitis prevention services in substance use disorders treatment clinics: a case study." Implement Sci **5(1): 46.**

Helfrich, C., Y.-F. Li, et al. (2009). "Organizational readiness to change assessment (ORCA): Development of an instrument based on the Promoting Action on Research in Health Services (PARIHS) framework." Implementation Science **4(1): 38.**

Helfrich, C. D., D. Blevins, et al. (2011). "Predicting implementation from organizational readiness for change: a study protocol." Implement Sci **6(1): 76.**

James, L. R. (1982). "Aggregation bias in estimates of perceptual agreement." Journal of Applied Psychology **67: 219-229.**

Klein, K. J. and S. W. J. Kozlowski (2000). Multilevel theory, research, and methods in organizations : foundations, extensions, and new directions. San Francisco, Jossey-Bass.

LeBreton, J. M. and J. L. Senter (2008). "Answers to 20 questions about interrater reliability and interrater agreement." Organizational Research Methods **11: 815-852.**

Shrout, P. E. and J. L. Fleiss (1979). "Intraclass Correlations: Uses in Assessing Rater Reliability." Psychological Bulletin **86(2): 420-428.**

ADDITIONAL SLIDES

Composition versus compilation forms of emergent constructs

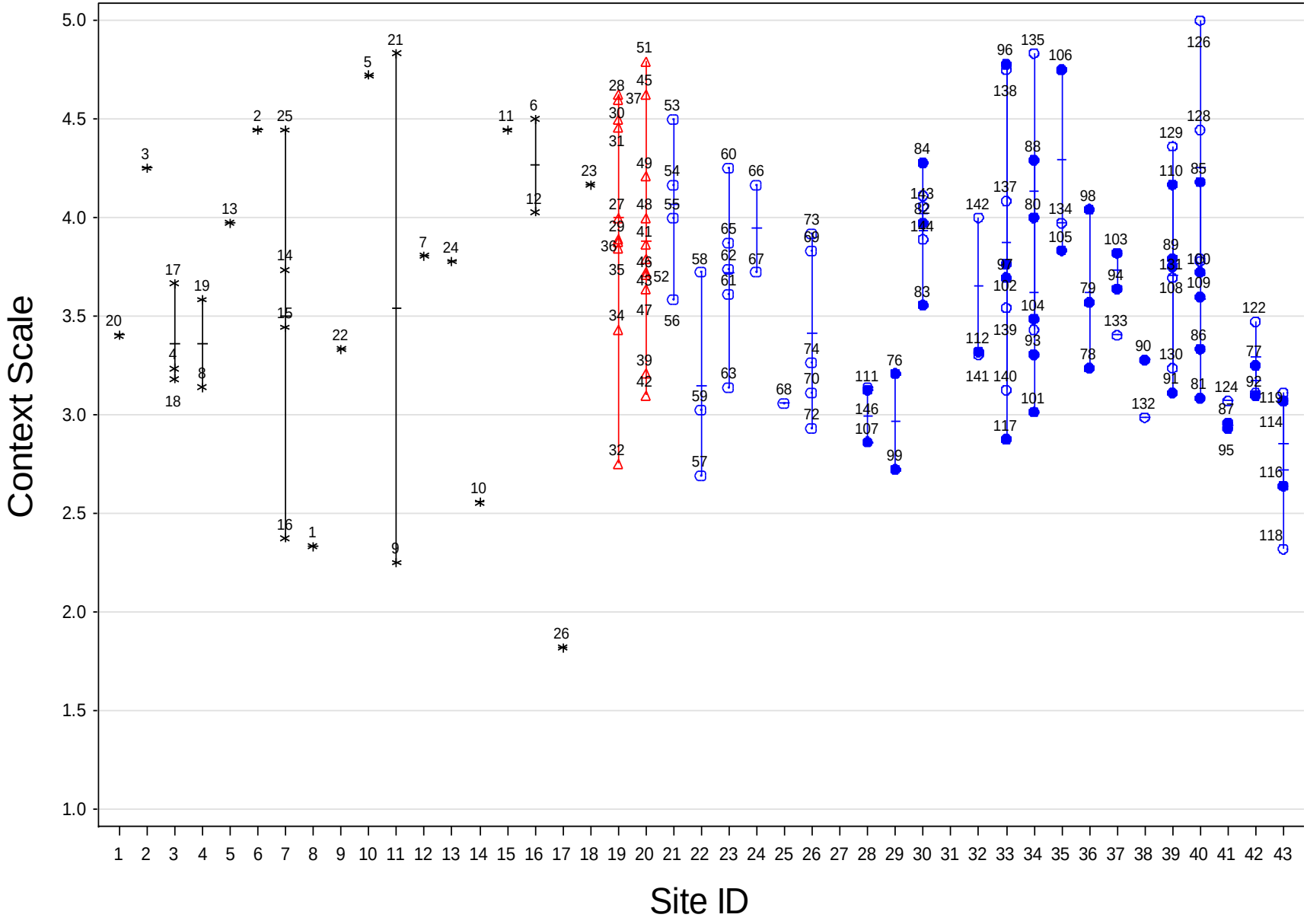
- IRR+IRA important for composition forms of emergence but not for compilation forms.
- Composition form of emergent construct is uniform at different levels
- Example of compositional construct: organizational climate for service, where we expect the experience of the organizational climate to be experienced in a similar way by all employees.
- Example of compilational construct: team performance.

Discussion: Comparison of ORCA IRA + IRR with organizational culture data

		Average *n* per site	ICC1	ICC2
ORCA	Evidence	2.3	**.32**	.52
	Context	2.7	**.27**	.48
CVF	Entrepreneurial	583	.03	**.95**
	Team	591	.03	**.95**
	Rational	590	.02	**.94**
	Hierarchical	588	.01	**.85**

CVF = competing values framework, a measure or organizational culture.

ORCA Scales: Values for each Rater, by PI and Site



Precision (reliability) vs. Accuracy (validity)

High accuracy (validity),
low precision (reliability)

Low accuracy (validity),
High precision (reliability)

- Source: Wikipedia, “Accuracy and Precision.” Accessed 10/23/11: http://en.wikipedia.org/wiki/Accuracy_and_precision